RISK SET SAMPLING DESIGNS FOR PROPORTIONAL HAZARDS MODELS

Ørnulf Borgan
Institute of Mathematics, University of Oslo,
P.O. Box 1053 Blindern, N-0316 Oslo, Norway

Bryan Langholz
Department of Preventive Medicine,
University of Southern California, School of Medicine,
1540 Alcazar Street, CHP-220, Los Angeles, California 90033, U.S.A.

June 1997

Abstract

The last five years, important progress has been made in understanding risk set sampling designs – like nested case-control studies – and in developing new and useful designs and statistical methods for sampled cohort data. The main purpose of this report is to present a fairly non-technical review of this development. We also illustrate the use of the general methodology in two particular situations: *(i)* a study on the effect of radon and smoking on the risk of lung cancer deaths among a cohort of uranium miners, and *(ii)* introduction of a "neighborhood-matched counter-matched" design particularly developed to investigated factors which may explain the observed association between "very high current configuration" power lines and childhood leukemia. The report will appear as a chapter in the book *Recent Advances in the Statistical Analysis of Medical Data* to be published by Edward Arnold Publishers Ltd. with Brian Everitt and Graham Dunn as editors.

RISK SET SAMPLING DESIGNS FOR PROPORTIONAL HAZARDS MODELS

Ørnulf Borgan and Bryan Langholz
University of Oslo and University of Southern California

1 Introduction

Cox's regression model (Cox, 1972) and similar proportional hazards models are central to modern survival analysis, and they are the methods of choice when one wants to assess the influence of risk factors and other covariates on mortality or morbidity. Estimation in such proportional hazards models is based on Cox's partial likelihood [see (2) below], which at each observed death or disease occurrence (failure) compares the covariate values of the failing individual to those of all individuals at risk at the time of the failure. In large epidemiologic cohort studies of a rare disease, (standard) use of proportional hazards models requires collection of covariate information on all individuals in the cohort even though only a small fraction of these actually get diseased. This may be very expensive, or even logistically impossible. Cohort sampling techniques, where covariate information is collected for all failing individuals (cases), but only for a sample of the non-failing individuals (controls) then offer useful alternatives which may drastically reduce the resources that need to be allocated to a study. Further, as most of the statistical information is contained in the cases, such studies may still be sufficient to give reliable answers to the questions of interest.

The most common cohort sampling design is nested case-control sampling, where for each case a small number of controls are selected at random from those at risk at the case's failure time, and where a new sample of controls is selected for each case. This risk set sampling technique was first suggested by Thomas (1977), who proposed to base inference on a modification of Cox's partial likelihood. This suggestion was supported by the work of Prentice and Breslow (1978), who derived the same expression as a conditional likelihood for time-matched case-control sampling from an infinite population. A more decisive, but still heuristic, argument was provided by Oakes (1981), who showed that one indeed gets a partial likelihood when the sampling of controls is performed within the actual finite cohort. It took more than ten years, however, before Goldstein and Langholz (1992) proved rigorously that the estimator of the regression coefficients based on Oakes' partial likelihood enjoys similar large sample properties as ordinary maximum likelihood estimators.

Goldstein and Langholz's paper initiated further work on risk set sampling methodology, and important progress has been achieved during the last few years both with respect to its theoretical foundation and the development of new methodology of practical importance. The key to this progress has been to model jointly the occurrence of failures and the sampling of controls as a marked point process (Borgan, Goldstein and Langholz, 1995). This marked point process formulation not only gives a more direct proof of Goldstein and Langholz's result. It also solves the problem of how to estimate the baseline hazard rate from nested case-control data, and it makes it simple to study other useful sampling schemes for the controls. In particular Borgan and Langholz (1993) discussed baseline hazard estimation for Cox's model for the relative mortality, while Langholz and Borgan (1995) studied a stratified version of nested case-control sampling, which they denoted counter matching, using this machinery. [Counter-matching was first proposed and studied by Langholz in a technical report using the approach of Goldstein and Langholz (1992).]

The purpose of this paper is to give a fairly non-technical review of this development. We do want to give the readers a flavor of the general theory, however, so we present heuristic arguments for many of our results – arguments which may be made rigorous using marked point processes, counting processes and martingales [see Borgan, Goldstein and Langholz (1995), for a detailed study Cox's regression model]. The outline of the paper is as follows. In Section 2 we first introduce a proportional hazards model with a general relative risk function and give some specific examples of such models including Cox's regression model. Then we describe the type of failure time data we consider for the cohort and remind the readers about the usual methods of inference for cohort data. In Section 3 the sampling of controls is discussed. We first review nested case-control sampling, and in particular point out how this design may be described by a uniform sampling distribution over the sets of potential controls. Then we describe in detail counter-matched (or stratified) sampling of the controls, and here as well we specify the sampling design in terms of its sampling distribution over sets of potential controls. Finally in this section, we introduce a general framework for the sampling of controls including nested case-control sampling and counter-matched sampling as special cases. In Section 4 we derive a partial likelihood, generalizing Oakes (1981) partial likelihood, for estimation of the regression coefficients. Section 5 is concerned with estimation of cumulative hazard rates from case-control data. We review how the cumulative baseline hazard rate may be estimated, and show how this forms the basis for estimation of cumulative hazard rates for individuals with given covariate histories. In Sections 2–5 we consider for simplicity proportional hazards models with a common baseline hazard rate for all individuals. In Section 6 this is relaxed by allowing the baseline hazard to differ between population strata. It is discussed how such stratified models are related to matching in epidemiologic studies, and how one at the analysis stage may "pool" baseline hazard estimates across population strata when a matched design has been used but turned out not to be really necessary. Sections 7 and 8 provide illustrations and extensions of the theory reviewed in earlier sections. In Section 7 the methods are illustrated by studying the effect of radon exposure and smoking on the risk of lung cancer deaths among a cohort of uranium miners from the Colorado Plateau, while in Section 8 we discuss a new design using neighborhood-matched counter-matching. In the final Section 9 we compare the application of the methods in our two examples.

2 Model and inference for cohort data

We consider a cohort of n individuals and denote by $\lambda_i(t) = \lambda(t; \mathbf{z}_i(t))$ the hazard rate at time t for an individual i with vector of covariates $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{ip}(t))'$. Here the time-variable t may be age, time since employment, or some other time-scale relevant to the problem at hand. The covariates may be time-fixed (like gender) or time-dependent (like cumulative exposure), and they may be indicators for categorical covariates (like the exposure groups "non-exposed," "low," "medium," and "high") or numeric (as when actual amount of exposure is recorded). We assume that the covariates of individual i is related to its hazard rate by the proportional hazards model

$$\lambda_i(t) = \lambda_0(t) r(\boldsymbol{\beta}, \mathbf{z}_i(t)). \tag{1}$$

Here $r(\boldsymbol{\beta}, \mathbf{z}_i(t))$ is a relative risk function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients describing the effect of the covariates, while the baseline hazard rate $\lambda_0(t)$ is left unspecified. We normalize the relative risk function by assuming $r(\boldsymbol{\beta}, \mathbf{0}) = 1$. Thus $\lambda_0(t)$ corresponds to the hazard rate of an individual with all covariates identically equal to zero. For the exponential relative risk function $r(\boldsymbol{\beta}, \mathbf{z}_i(t)) = \exp(\boldsymbol{\beta}' \mathbf{z}_i(t))$, formula (1) gives the usual Cox regression model.

Other possibilities include the linear relative risk function $r(\boldsymbol{\beta}, \mathbf{z}_i(t)) = 1+\boldsymbol{\beta}'\mathbf{z}_i(t)$ and the excess relative risk model $r(\boldsymbol{\beta}, \mathbf{z}_i(t)) = \prod_{j=1}^{p}(1 + \beta_j z_{ij}(t))$. Even though it is not made explicit in our notation, we will also allow for an "offset" in the model, i.e., a covariate for which no regression parameter is estimated. One such example is Cox's regression model for the relative mortality (Andersen *et al.*, 1985; Borgan and Langholz, 1993).

The individuals in the cohort may be followed over different periods of time, i.e., our observations may be subject to left-truncation and/or right censoring. The risk set $\mathcal{R}(t)$ is the collection of all individuals who are under observation just before time t, and $n(t) = |\mathcal{R}(t)|$ is the number at risk at that time. We let $t_1 < t_2 < \cdots$ be the times when failures are observed and, assuming that there are no tied failures, denote by i_j the index of the individual who fails at t_j (a few ties may be broken at random). We assume throughout that truncation and censoring are independent in the sense that the additional knowledge of which individuals have entered the study or have been censored before any time t do not carry information on the risks of failure at t (see Sections III.2-3 in Andersen, Borgan, Gill and Keiding, 1993, for a general discussion). Then the vector of regression parameters in (1) is estimated by $\widehat{\boldsymbol{\beta}}$, the value of $\boldsymbol{\beta}$ maximizing Cox's partial likelihood

$$L_{\mathrm{c}}(\boldsymbol{\beta}) = \prod_{t_j} \frac{r(\boldsymbol{\beta}, \mathbf{z}_{i_j}(t_j))}{\sum_{l \in \mathcal{R}(t_j)} r(\boldsymbol{\beta}, \mathbf{z}_l(t_j))}. \tag{2}$$

The cumulative baseline hazard rate $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$ is estimated by the Breslow estimator

$$\widehat{\Lambda}_0(t) = \sum_{t_j \le t} \frac{1}{\sum_{l \in \mathcal{R}(t_j)} r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_l(t_j))}. \tag{3}$$

It is well known that $\widehat{\boldsymbol{\beta}}$ enjoys similar large sample properties as an ordinary maximum likelihood estimator, while the Breslow estimator (properly normalized) asymptotically is distributed as a Gaussian process (e.g. Andersen, Borgan, Gill and Keiding, 1993, Section VII.2). In particular $\widehat{\Lambda}_0(t)$ is asymptotically normally distributed for any given value of t.

3 Sampling of controls

From (2) and (3) it is seen, as already indicated in the introduction, that covariate information is needed for all individuals at risk in order to apply the usual inference methods for cohort data. A similar methodology is available when we only have covariate information for the failing individuals (cases) and control individuals sampled from those at risk at the times of the failures. We will review this methodology in Sections 4 and 5. But before we do that, we need to describe more precisely how the controls are selected. We will first consider the nested case-control design and counter-matched sampling, which are the two most important risk set sampling techniques. Then we will describe a general framework for the sampling of controls which contain these two as special cases.

3.1 Nested case-control sampling

Consider the "classical" nested case-control design due to Thomas (1977). Here, if an individual i fails at time t, one selects $m-1$ controls by simple random sampling (without replacement) from the $n(t)-1$ non-failing individuals in the risk set $\mathcal{R}(t)$. The set $\widetilde{\mathcal{R}}(t)$ consisting of the case i and these $m-1$ controls is denoted the sampled risk set. Note that sampling is done

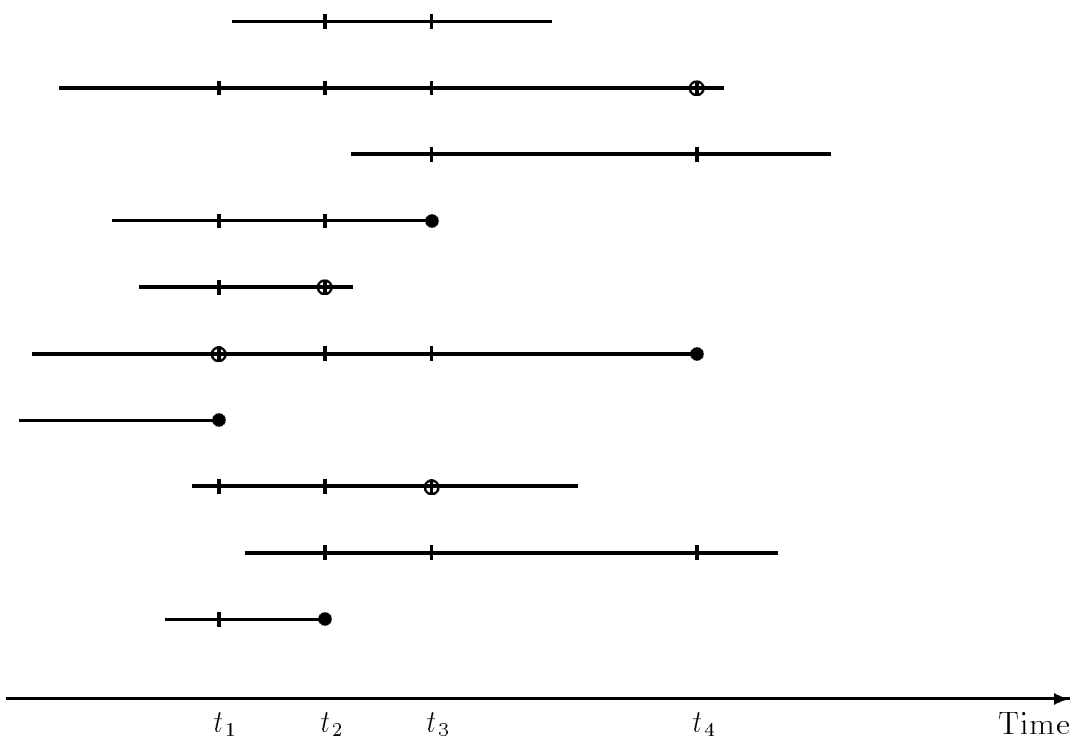

Figure 1: Illustration of risk set sampling, with one control per case, from a hypothetical cohort of 10 individuals. Each individual is represented by a line starting at an entry time and ending at an exit time corresponding to censoring or failure. Failure times are indicated by dots (•), non-failing individuals at risk at the failure times are indicated by bars (|) and the sampled controls are indicated by circles (○).

independently across risk sets so that subjects may serve as controls for multiple cases and cases may serve as controls for other cases that failed when the case was at risk.

Figure 1 illustrates the basic features of a nested case-control study for a small hypothetical cohort of 10 individuals with one control selected per case (i.e. $m = 2$). Each individual in the cohort is represented by a horizontal line starting at some entry time and ending at some exit time. If the exit time corresponds to a failure, this is represented by a "•" in the figure. In the hypothetical cohort considered, four individuals are observed to fail. The potential controls for these four cases are indicated by a "|" in the figure, and are given as all non-failing individuals at risk at the times of the failures. Among the potential controls one is selected at random as indicated by a "○" in the figure. The four sampled risk sets are then represented by the four •, ○ pairs in Figure 1.

In what follows we will not only need to know which individuals were actually selected as controls. For the inference procedures discussed in Sections 4 and 5, it is crucial also to know the probability of selecting certain sets of individuals as our controls. It turns out to be convenient to describe the sampling scheme for the controls by the conditional probability of selecting a given set $\mathbf{r}$ as our sampled risk set $\tilde{\mathcal{R}}(t)$, given that an individual i fails at time t and given the risk set $\mathcal{R}(t)$. Since the $m-1$ controls are selected at random among the $n(t)-1$ non-failing

individuals at risk, we have

$$\Pr\left(\tilde{\mathcal{R}}(t) = \mathbf{r} \mid i \text{ fails at } t,\, \mathcal{R}(t)\right) = \binom{n(t)-1}{m-1}^{-1} = \binom{n(t)}{m}^{-1} \frac{n(t)}{m} \tag{4}$$

for any set $\mathbf{r} \subset \mathcal{R}(t)$ which contains i and is of size $|\mathbf{r}| = m$. Here the last equality follows since $\binom{n(t)}{m} = \binom{n(t)-1}{m-1}\frac{n(t)}{m}$. Note that the right-most expression in (4) gives a factorization into a probability distribution

$$\pi_t(\mathbf{r}) = \binom{n(t)}{m}^{-1} I(|\mathbf{r}| = m) \tag{5}$$

over sets $\mathbf{r} \subset \mathcal{R}(t)$ and a weight

$$w_i(t) = \frac{n(t)}{m} I(i \in \mathbf{r}). \tag{6}$$

This factorization will be useful below.

3.2 Counter-matching

To select a nested case-control sample, only the at risk status of the individuals in the cohort is needed. Often, however, some additional information is available for all cohort members, e.g., a surrogate measure of exposure, like type of work or duration of employment, may be available for everyone. Langholz and Borgan (1995) have developed a stratified version of the nested case-control design which makes it possible to incorporate such information into the sampling process in order to obtain a more informative sample of controls. For this design, called counter-matching, one applies the additional information on the cohort subjects to classify each individual at risk into one of say, L, strata. We denote by $\mathcal{R}_l(t)$ the subset of the risk set $\mathcal{R}(t)$ which belongs to stratum l, and let $n_l(t) = |\mathcal{R}_l(t)|$ be the number at risk in this stratum just before time t. If a failure occurs at t, we want to sample our controls such that the sampled risk set will contain m_l individuals from each stratum $l = 1, \ldots, L$. This is obtained as follows. Assume that an individual i who belongs to stratum $s(i)$ fails at t. Then for $l \neq s(i)$ one samples randomly without replacement m_l controls from $\mathcal{R}_l(t)$. From the case's stratum $s(i)$ only $m_{s(i)} - 1$ controls are sampled. The failing individual i is, however, included in the sampled risk set $\tilde{\mathcal{R}}(t)$, so this contains a total of m_l from each stratum. Even though it is not made explicit in the notation, we note that the classification into strata may be time-dependent. A crucial assumption, however, is that the information on which the stratification is based has to be known just before time t.

In probabilistic terms, counter-matched sampling may be described as follows. For any given set $\mathbf{r} \subset \mathcal{R}(t)$ which contains i and satisfies $|\mathbf{r} \cap \mathcal{R}_l(t)| = m_l$ for $l = 1, \ldots, L$, we have

$$\begin{aligned} &\Pr\left(\tilde{\mathcal{R}}(t) = \mathbf{r} \mid i \text{ fails at } t,\, \mathcal{R}_l(t); l = 1, \ldots, L\right) \\ &= \left\{ \binom{n_{s(i)}(t)-1}{m_{s(i)}-1} \prod_{l \neq s(i)} \binom{n_l(t)}{m_l} \right\}^{-1} = \left\{ \prod_{l=1}^{L} \binom{n_l(t)}{m_l} \right\}^{-1} \frac{n_{s(i)}(t)}{m_{s(i)}}. \end{aligned} \tag{7}$$

Note that the last expression gives a factorization into a probability distribution

$$\pi_t(\mathbf{r}) = \left\{ \prod_{l=1}^{L} \binom{n_l(t)}{m_l} \right\}^{-1} I(|\mathbf{r} \cap \mathcal{R}(t)| = m_l;\, l = 1, \ldots, L) \tag{8}$$

over sets $\mathbf{r} \subset \mathcal{R}(t)$ and a weight

$$w_i(t) = \frac{n_{s(i)}(t)}{m_{s(i)}} I(i \in \mathbf{r}). \tag{9}$$

By counter-matching, one may be able to increase the variation in the value of the covariate of main interest within each sampled risk set, and this will increase the statistical efficiency for estimating the corresponding regression coefficient. In particular, if this covariate is binary, and we select one control per case, concordant pairs (i.e., the case and its control have the same value of the covariate) do not give any information in estimating the effect of the covariate. For a counter-matched design with $L = 2$ and $m_1 = m_2 = 1$, and where stratification is based on a surrogate correlated with the covariate of interest, the single control is selected from the opposite stratum of the case. This will reduce the number of concordant pairs, and thereby increase the information contained in the matched pairs of cases and controls. The situation with two strata and one control per case also gives a motivation for the name counter-matching. As the name suggests, it is essentially the opposite of matching where the case and its control are from the same stratum (cf. Section 6).

3.3 General sampling designs

In order to describe a general model for the sampling of controls, we first need to introduce "the history" $\mathcal{F}_{t-}$, which contains information about events (entries, exits, changes in covariate values) in the cohort as well as on the sampling of controls, up to, but not including, time t. Only part of this information, like the numbers at risk in different strata, will be available to the researcher in a case-control study. Based on the information which actually is available just before time t, one decides on a sampling strategy for the controls. This may be described in probabilistic terms as follows. If an individual i fails at time t, the set $\mathbf{r} \subset \mathcal{R}(t)$ is selected as our sampled risk set $\widetilde{\mathcal{R}}(t)$ with probability $\pi_t(\mathbf{r} \mid i)$. The sampled risk set consists of the case and its controls, so we let $\pi_t(\mathbf{r} \mid i) = 0$ when $i \notin \mathbf{r}$. With this convention $\pi_t(\cdot \mid i)$ is a probability distribution over all sets $\mathbf{r} \subset \mathcal{R}(t)$. Note that for nested case-control sampling and counter-matched sampling, $\pi_t(\mathbf{r} \mid i)$ is given by (4) and (7), respectively. The full cohort study is also a special case of this general framework in which the full risk set is sampled with probability one, i.e., $\pi_t(\mathbf{r} \mid i) = I(\mathbf{r} = \mathcal{R}(t))$ for all $i \in \mathcal{R}(t)$. Other designs (quota sampling and counter-matching with additionally randomly sampled controls) are discussed by Borgan, Goldstein and Langholz (1995) and Langholz and Goldstein (1996). We note that the sampling of controls may depend in an arbitrary way on events in the past (which are known to the researcher), i.e., on events which are contained in $\mathcal{F}_{t-}$. It may, however, not depend on events in the future. For example, one may not exclude as potential controls for a current case individuals that subsequently fail.

In connection with (4) and (7), we introduced a factorization of the relevant $\pi_t(\mathbf{r} \mid i)$ into a sampling distribution $\pi_t(\mathbf{r})$ over sets $\mathbf{r} \subset \mathcal{R}(t)$ and a weight $w_i(t)$. A similar factorization is possible for the general case as well. To this end we introduce

$$\pi_t(\mathbf{r}) = n(t)^{-1} \sum_{l=1}^{n} \pi_t(\mathbf{r} \mid l), \tag{10}$$

which is a probability distribution over all sets $\mathbf{r} \subset \mathcal{R}(t)$. The formulas (5) and (8) are special cases of (10). We also introduce the weights

$$w_i(t) = \frac{\pi_t(\mathbf{r} \mid i)}{n(t)^{-1} \sum_{l=1}^{n} \pi_t(\mathbf{r} \mid l)}, \tag{11}$$

and note that (6) and (9) are special cases of this formula. [It should be realized that the general weights (11), as well as the special cases (6) and (9), do depend on the set $\mathbf{r}$. We have, however, chosen not to make this explicit in the notation.] Corresponding to (4) and (7), we then have the factorization

$$\Pr\left(\widetilde{\mathcal{R}}(t) = \mathbf{r} \mid i \text{ fails at } t,\, \mathcal{F}_{t-}\right) = \pi_t(\mathbf{r} \mid i) = \pi_t(\mathbf{r}) w_i(t) \tag{12}$$

for sets $\mathbf{r} \subset \mathcal{R}(t)$.

For all the sampling designs, we assume that the selection of controls is done independently at the different failure times, so that an individual may be a member of more than one sampled risk set. Further, a basic assumption throughout is that not only the truncation and censoring, but also the sampling of controls, are independent in the sense that the additional knowledge of which individuals have entered the study, have been censored or have been selected as controls before any time t do not carry information on the risks of failure at t. This assumption will be violated if, e.g., in a prevention trial, individuals selected as controls change their behavior in such a way that their risk of failure is different from similar individuals who have not been selected as controls. If we introduce $[\mathrm{d}t)$ as a short-hand notation for the small time-interval $[t, t + \mathrm{d}t)$, the above independence assumption and (1) imply that

$$\Pr\left(i \text{ fails in } [\mathrm{d}t) \mid \mathcal{F}_{t-}\right) = r(\boldsymbol{\beta}, \mathbf{z}_i(t)) \lambda_0(t) \mathrm{d}t \tag{13}$$

when $i \in \mathcal{R}(t)$.

4 Partial likelihood and estimation of the regression coefficients

Estimation of the regression coefficients in (1) is based on a partial likelihood which may be derived in a similar manner as Cox's partial likelihood (2) for the full cohort. Heuristically the argument goes as follows. Consider a set $\mathbf{r} \subset \mathcal{R}(t)$ and an individual $i \in \mathbf{r}$. Then by (12) and (13)

$$\begin{aligned} &\Pr\left(i \text{ fails in } [\mathrm{d}t),\, \widetilde{\mathcal{R}}(t) = \mathbf{r} \mid \mathcal{F}_{t-}\right) \\ &\quad = \Pr\left(i \text{ fails in } [\mathrm{d}t) \mid \mathcal{F}_{t-}\right) \times \Pr\left(\widetilde{\mathcal{R}}(t) = \mathbf{r} \mid i \text{ fails at } t,\, \mathcal{F}_{t-}\right) \\ &\quad = r(\boldsymbol{\beta}, \mathbf{z}_i(t)) \lambda_0(t) \mathrm{d}t \times \pi_t(\mathbf{r} \mid i) = r(\boldsymbol{\beta}, \mathbf{z}_i(t)) w_i(t) \pi_t(\mathbf{r}) \lambda_0(t) \mathrm{d}t. \end{aligned} \tag{14}$$

Now the sampled risk set equals $\mathbf{r}$ if one of the individuals in $\mathbf{r}$ fails, and the remaining ones are selected as controls. Therefore

$$\Pr\left(\text{one failure in } \mathbf{r} \text{ in } [\mathrm{d}t),\, \widetilde{\mathcal{R}}(t) = \mathbf{r} \mid \mathcal{F}_{t-}\right) = \sum_{l \in \mathbf{r}} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t) \pi_t(\mathbf{r}) \lambda_0(t) \mathrm{d}t. \tag{15}$$

Dividing (14) by (15), it follows that

$$\Pr\left(i \text{ fails at } t \mid \text{ one failure in } \mathbf{r} \text{ at } t,\, \widetilde{\mathcal{R}}(t) = \mathbf{r},\, \mathcal{F}_{t-}\right) = \frac{r(\boldsymbol{\beta}, \mathbf{z}_i(t)) w_i(t)}{\sum_{l \in \mathbf{r}} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t)}. \tag{16}$$

We then multiply together conditional probabilities of the form (16) for all observed failure times t_j, cases i_j, and sampled risk sets $\widetilde{\mathcal{R}}(t_j)$, and obtain the partial likelihood

$$L_{\mathrm{s}}(\boldsymbol{\beta}) = \prod_{t_j} \frac{r(\boldsymbol{\beta}, \mathbf{z}_{i_j}(t)) w_{i_j}(t_j)}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t_j)}. \tag{17}$$

This is similar to the full cohort partial likelihood (2), except that the sum in the denominator only is taken over the sampled risk set $\widetilde{\mathcal{R}}(t_j)$ and that the contribution of each individual (including the case) has to be weighted by $w_l(t_j)$ to compensate for the differences in the sampling probabilities. In fact, (2) is the special case of (17) in which the entire risk set is sampled with probability one and all weights are unity. Inference concerning $\boldsymbol{\beta}$, using the usual large sample likelihood methods, can be based on the partial likelihood (17). In particular the maximum partial estimator $\widehat{\boldsymbol{\beta}}$ is approximately multinormally distributed around the true parameter vector $\boldsymbol{\beta}$ with a covariance matrix that may be estimated as $\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}$, the inverse of the observed information matrix. Formal proofs, along the lines of Andersen and Gill (1982), are provided by Borgan, Goldstein and Langholz (1995) for Cox's regression model.

Note that for nested case-control sampling, the weights (6) are the same for all individuals and hence cancel from (17) giving Oakes (1981) partial likelihood. In fact, the above heuristic derivation of (17) is parallel to the one originally given by Oakes for simple random sampling of the controls. Borgan, Goldstein and Langholz (1995) made this argument rigorous and extended it to general sampling designs using a marked point processes formulation.

When we have an exponential relative risk function $r(\boldsymbol{\beta}, \mathbf{z}_i(t)) = \exp(\boldsymbol{\beta}' \mathbf{z}_i(t))$, the partial likelihood (17) is formally the same as a weighted conditional logistic regression likelihood used in the analysis of matched case-control studies. Standard software packages which have modules for the analysis matched case-control studies, such as SAS PHREG, EGRET, EPILOG, or EPICURE, may therefore be used to estimate $\boldsymbol{\beta}$. The weights are accommodated by including the weight as a covariate and fixing the parameter associated with it to one. The package EPICURE fits a wide variety of relative risk functions $r(\boldsymbol{\beta}, \mathbf{z}_i(t))$ and was used to estimate parameters from the Colorado Plateau uranium miners data in Section 7.

5 Estimation of cumulative hazard rates

The cumulative baseline hazard rate $\Lambda_0(t) = \int_0^t \lambda_0(u) \mathrm{d}u$ can be estimated by

$$\widehat{\Lambda}_0(t; \widehat{\boldsymbol{\beta}}) = \sum_{t_j \le t} \frac{1}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_l(t_j)) w_l(t_j)}. \tag{18}$$

The estimator (18) is of the same form as the Breslow estimator (3) for cohort data, but with the same modifications as for the partial likelihood (17). Here as well the full cohort estimator is obtained as the special case where the entire risk set is sampled with probability one and all weights are unity. The estimator (18) was first introduced by Borgan and Langholz (1993) for nested case-control studies in the context of Cox's model for the relative mortality. Borgan, Goldstein and Langholz (1995) considered general sampling designs and studied the large sample properties of (18) for Cox's regression model using theory for counting processes, martingales and stochastic integrals.

The following heuristic argument gives a motivation for the estimator (18). Consider the increment over $[\mathrm{d}t)$ of $\widehat{\Lambda}_0(t; \boldsymbol{\beta})$ defined as in (18), but with $\widehat{\boldsymbol{\beta}}$ replaced by the true value $\boldsymbol{\beta}$. This

increment equals

$$1 \Big/ \sum_{l \in \widetilde{\mathcal{R}}(t)} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t) \tag{19}$$

if a failure occurs at t and the sampled risk set is $\widetilde{\mathcal{R}}(t)$, and it is zero otherwise. By (15), and since (10) is a probability distribution over sets $\mathbf{r} \subset \mathcal{R}(t)$, it follows that, given $\mathcal{F}_{t-}$, the expected value of the increment is

$$\begin{aligned}
&\sum_{\mathbf{r} \subset \mathcal{R}(t)} \frac{1}{\sum_{l \in \mathbf{r}} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t)} \times \Pr\left(\text{one failure in } \mathbf{r} \text{ in } [\mathrm{d}t),\ \widetilde{\mathcal{R}}(t) = \mathbf{r} \,|\, \mathcal{F}_{t-}\right) \\
&\quad = \sum_{\mathbf{r} \subset \mathcal{R}(t)} \frac{\sum_{l \in \mathbf{r}} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t) \pi_t(\mathbf{r}) \lambda_0(t) \mathrm{d}t}{\sum_{l \in \mathbf{r}} r(\boldsymbol{\beta}, \mathbf{z}_l(t)) w_l(t)} \\
&\quad = \sum_{\mathbf{r} \subset \mathcal{R}(t)} \pi_t(\mathbf{r}) \lambda_0(t) \mathrm{d}t \ = \ \lambda_0(t) \mathrm{d}t,
\end{aligned}$$

i.e., the increment of $\Lambda_0(t)$ over $[\mathrm{d}t)$. Thus (18) is almost unbiased when averaged over all possible failure and sampled risk set occurrences, see Borgan, Goldstein and Langholz (1995) for a rigorous argument using martingales.

Let us then consider estimation of the cumulative hazard rate

$$\Lambda(t; \mathbf{z}_0) = \int_0^t r(\boldsymbol{\beta}_0; \mathbf{z}_0(u)) \lambda_0(u) \mathrm{d}u = \int_0^t r(\boldsymbol{\beta}_0; \mathbf{z}_0(u) \mathrm{d}\Lambda_0(u)$$

corresponding to an individual with a specified time-dependent covariate history $\mathbf{z}_0(u); 0 < u \leq t$. By (19) this may be estimated by

$$\widehat{\Lambda}(t; \mathbf{z}_0) = \sum_{t_j \leq t} \frac{r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0(t_j))}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_l(t_j)) w_l(t_j)}. \tag{20}$$

In order to estimate the variance of (20), we introduce

$$\widehat{\omega}^2(t; \mathbf{z}_0) = \sum_{t_j \leq t} \left\{ \frac{r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0(t_j))}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}}, \mathbf{z}_l(t_j)) w_l(t_j)} \right\}^2,$$

and

$$\widehat{\mathbf{B}}(t; \mathbf{z}_0) = \sum_{t_j \leq t} \left\{ \frac{\dot{\mathbf{r}}(\widehat{\boldsymbol{\beta}}; \mathbf{z}_0(t_j))}{\sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}}; \mathbf{z}_l(t_j)) w_l(t_j)} - \frac{r(\widehat{\boldsymbol{\beta}}; \mathbf{z}_0(t_j)) \sum_{l \in \widetilde{\mathcal{R}}(t_j)} \dot{\mathbf{r}}(\widehat{\boldsymbol{\beta}}; \mathbf{z}_l(t_j)) w_l(t_j)}{\left\{ \sum_{l \in \widetilde{\mathcal{R}}(t_j)} r(\widehat{\boldsymbol{\beta}}; \mathbf{z}_l(t_j)) w_l(t_j) \right\}^2} \right\}$$

with

$$\dot{\mathbf{r}}(\boldsymbol{\beta}; \mathbf{z}(u)) = \frac{\partial}{\partial \boldsymbol{\beta}} r(\boldsymbol{\beta}; \mathbf{z}(u)).$$

Then $\widehat{\Lambda}(t;\mathbf{z}_0)$ is asymptotically normally distributed around its true value $\Lambda(t;\mathbf{z}_0)$ with a variance that may be estimated by

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(t;\mathbf{z}_0)) = \widehat{\omega}^2(t;\mathbf{z}_0) + \widehat{\mathbf{B}}(t;\mathbf{z}_0)'\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\widehat{\mathbf{B}}(t;\mathbf{z}_0) \tag{21}$$

(Langholz and Borgan, 1997). Here the leading term on the right hand side is due to the variability in estimating the hazard while the second term accounts for the variability due to the estimation of the relative risk parameters $\boldsymbol{\beta}$. Note that the variance estimator for the cumulative baseline hazard rate estimator (18) is the special case of (21) obtained by letting $\mathbf{z}_0(t) = \mathbf{0}$ for all t. Note also that if the cumulative hazard rate $\Lambda(s,t;\mathbf{z}_0) = \int_s^t r(\boldsymbol{\beta}_0;\mathbf{z}_0(u))\lambda_0(u)\mathrm{d}u$ over the interval $(s,t]$ is to be estimated, the above formulas still apply provided that the sums are restricted to the failure times t_j falling in this interval.

6 Matching and pooling

In order to keep the presentation simple, we have so far considered the proportional hazards model (1) where the baseline hazard rate is assumed to be the same for all individuals in the cohort. Sometimes this may not be reasonable, e.g., to control for the effect of one or more confounding factors, one may want to adopt a stratified version of (1) where the baseline hazard differ between (possibly time-dependent) population strata generated by the confounders. The regression coefficients are, however, assumed the same across these strata. Thus the hazard rate of an individual i from population stratum c is assumed to take the form

$$\lambda_i(t) = \lambda_{0c}(t)r(\boldsymbol{\beta},\mathbf{z}_i(t)). \tag{22}$$

When the stratified proportional hazards model (22) applies, the sampling of controls should be restricted to those at risk in the same population stratum as the case. We say that the controls are matched by the stratification variable. In particular for a matched nested case-control study, if an individual in population stratum c fails at time t, one selects at random $m-1$ controls from the $n_c(t)-1$ non-failing individuals at risk in this population stratum. Similarly one may combine matching and counter-matching by selecting the controls among those in the sampling strata used for counter-matching which belong to the population stratum of the case.[1] In general one obtains a matched case-control study by restricting the sampling distributions to those which only give positive probability to sets contained in the population stratum of the case.

The general theory of Sections 3–5 goes through almost unchanged for matched case-control sampling within the framework of the stratified proportional hazards model (22) provided one uses the sampling distribution $\pi_i(\mathbf{r})$ and weights $w_i(t_j)$ relevant to the sampling design. For sets $\mathbf{r}$ and individuals i in population stratum c these may be obtained from (10) and (11) if we replace $n(t)$ by $n_c(t)$, the number at risk in population stratum c just before time t, and restrict the sums to those individuals l who belong to this population stratum. In particular for a matched case-control study using a counter-matched design for control selection, the proper weights are $w_i(t_j) = n_{s(i),c}(t_j)/m_{s(i)}$. Here $s(i)$ denotes the sampling stratum of individual i, while $n_{s(i),c}(t_j)$ is the number of individuals at risk in sampling stratum $s(i)$ who also belong to the population stratum c of the case.

It follows that the partial likelihood (17) applies without modification for a matched case-control study provided one uses the proper weights as just described. Further, when there is

[1] It is important to distinguish between the population strata which forms the basis for stratification in (22) and the sampling strata used for counter-matched sampling of the controls. This distinction will be illustrated for the uranium miners data in the following section. There the population strata will correspond to different calendar periods, while counter-matching will be based on cumulative radon exposure.

only a small number of strata, the stratum specific cumulative baseline hazard rates $\Lambda_{0c}(t) = \int_0^t \lambda_{0c}(u)\mathrm{d}u$ may be estimated using these weights by a slight modification of (18). All that is required is that the sum is restricted to those failure times t_j when a failure in the actual population stratum occurs. When there are many population strata, however, there may be too little information in each stratum to make estimation of the stratum specific cumulative baseline hazard rates meaningful.

If the estimates for the stratum specific cumulative baseline hazard rates turn out to be quite similar (so that matching was not really necessary in the first place), one may at the analysis stage want to "pool" over the population strata to get a common estimator for the cumulative baseline hazard rate. Such a procedure is a special case of the results of Section 5. For assuming the model (1) with a common baseline hazard across the population strata, the general theory applies with the sampling probability distribution (12) giving zero probability to all sets $\mathbf{r}$ not contained in the population stratum of the case. It follows that the common cumulative baseline hazard may be estimated by (18), with variance estimator obtained from (21), using weights $w_i(t_j)$ which for population stratum c equal those used in the matched analysis times $n(t_j)/n_c(t_j)$. In particular for counter-matched sampling of the controls within each population stratum, the weights equal $w_i(t_j) = (n_{s(i),c}(t_j)/m_{s(i)}) \times (n(t_j)/n_c(t_j))$.

7 Lung cancer deaths among uranium miners

Our first illustration uses data from a cohort of uranium miners from the Colorado Plateau and repeats to some extent material earlier published by Langholz and Goldstein (1996) and Langholz and Borgan (1997).

7.1 Data and model

The Colorado Plateau uranium miners cohort was assembled to study the effects of radon exposure and smoking on mortality rates and has been described in detail in earlier publications (e.g. Lundin, Wagoner, and Archer, 1971; Hornung and Meinhardt, 1987). We will focus on lung cancer mortality. The cohort consists of 3,347 Caucasian male miners recruited between 1950 and 1960 and was traced for mortality outcomes through December 31, 1982, by which time 258 lung cancer deaths were observed. Exposure data included radon exposure, in working level months (WLM) (Committee on the Biological Effects of Ionizing Radiation, 1988, p. 27), and smoking histories, in number of packs of cigarettes (20 cigarettes per pack) smoked per day.

We consider age as the basic time scale and, as there has been a well known secular trend in lung cancer rates in the general United States population, calendar time was treated as a matching factor with levels defined as the six five year periods 1950-1954, 1955-1959,...,1975-1979, and 1980-1982. Although covariate information is available on all cohort subjects, in order to illustrate the methods we selected nested case-control and counter-matched samples with one and three controls per case from the risk sets formed by the case's age and his five-year calendar period at death. These data sets are denoted 1:1 and 1:3 nested case-control and counter-matched samples, respectively. The 23 tied failure times were broken randomly so that there was only one case per risk set. Following Langholz and Goldstein (1996), counter-matching was based on radon exposure grouped into two or four strata according to the quartiles of the cumulative radon exposure for the cases, and one control was sampled at random from each stratum except the one of the case. Details are provided in Section 5 in the paper by Langholz and Goldstein; cf. in particular their Table 2. Such a counter-matched design is useful for situations where exposure data (here radon) are available for everyone, while confounder information (here smoking) has to be collected from the case-control data, the goal being to

Table 1: Estimated regression coefficients (with standard errors) per 100 WLMs cumulative radon exposure and per 1000 packs of cigarettes smoked for the stratified excess relative risk model $\lambda(t) = \lambda_{0c}(t)(1 + \beta_R R(t))(1 + \beta_S S(t))$ for various risk set sampling designs.

Sampling design	Radon (β_R)	Smoke (β_S)
1:1 nested case-control	0.42 (0.20)	0.23 (0.10)
1:1 counter-matched	0.39 (0.14)	0.25 (0.10)
1:3 nested case-control	0.43 (0.16)	0.20 (0.07)
1:3 counter-matched	0.41 (0.13)	0.19 (0.07)
Cohort	0.38 (0.11)	0.17 (0.05)

assess the effect of the exposure after controlling for the confounder. If data on the exposure of main interest (here radon) are not available for everyone, one option is to counter-match on duration of employment used as a surrogate for exposure, and then collect precise exposure and confounder information for the sampled data (Steenland and Deddens, 1997).

We summarized radon and smoking data into cumulative exposure up to two years prior to the age of death of the case. Thus, we consider as covariates $\mathbf{z}(t) = (R(t), S(t))$, where $R(t)$ is cumulative radon exposure measured in working level months (WLM) up to two years prior to age t, and $S(t)$ is cumulative smoking in number of packs smoked up to two years prior to t. As has been the case in previous analyzes of these data (Whittemore and McMillan, 1983; Lubin *et al.*, 1994; Thomas *et al.*, 1994), the excess relative risk model was used. Thus the hazard rate is assumed to take the form $\lambda(t) = \lambda_{0c}(t)(1 + \beta_R R(t))(1 + \beta_S S(t))$ for the cth calendar period; cf. (22).

7.2 Relative and absolute risk

The regression parameter estimates are given in Table 1 for the different sampling designs. It is seen that both radon and smoking have a significant effect on the risk of lung cancer death when adjusted for the effect of the other, and that both the nested case-control designs and the counter-matched designs give estimates quite close to those obtained from the full cohort. The radon excess relative risk is about 0.4 per 100 WLMs cumulative radon exposure for all designs, while the smoking excess relative risk is about 0.2 per 1000 packs of cigarettes smoked. As expected the 1:1 nested case-control design has the largest standard errors, about twice the size of those from the full cohort. Counter-matching gives a substantial improvement in the precision of the estimates of the radon excess relative risk, e.g., the 1:1 counter-matched sample gives a more precise estimate of β_R than the 1:3 sample with simple random sampling of the controls. Usually counter-matching will reduce the precision of estimates of parameters of less importance (Langholz and Borgan, 1995; Langholz and Goldstein, 1996). Here, however, the estimates of β_S based on simple and stratified sampling of the controls have the same precision. This may be due to the commonness of smoking in the cohort and the fact that it is relatively uncorrelated to radon (Langholz and Clayton, 1994, Table 4).

We first estimated the cumulative baseline hazard separately for each calendar period. As these turned out to be quite similar, we decided to pool the estimates as described in Section 6 to get an estimate for the cumulative baseline hazard valid for all calendar periods. These "pooled" estimates are shown in Figure 2 for cohort data and the four case-control data sets.

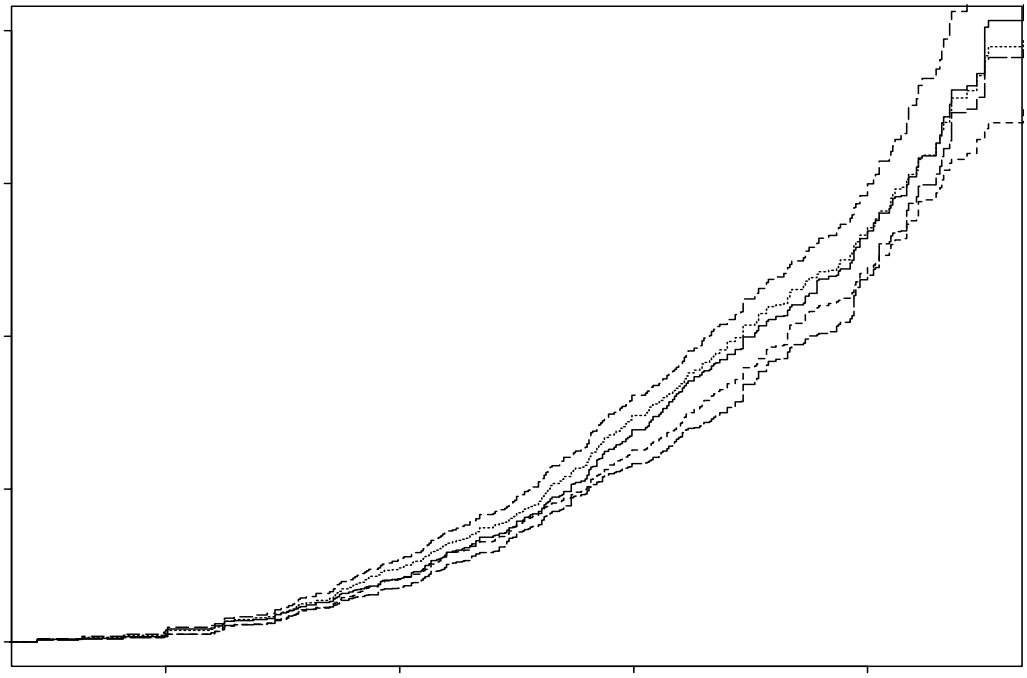

Figure 2: Estimated cumulative baseline hazard $\int_0^t \lambda_0(u)\mathrm{d}u$ of lung cancer deaths as a function of age from the excess relative risk model $\lambda(t) = \lambda_0(t)(1 + \beta_R R(t))(1 + \beta_S S(t))$ with common baseline hazard for all calendar periods (given from the lowest to the highest at age 60 years): 1:1 nested case-control (— — — —); 1:1 counter-matched (- - - - - - -); 1:3 nested case-control (————————-); 1:3 counter-matched (· · · · · · · · · · · ·); cohort (– – – – –).

The sampled data give cumulative baseline hazard estimates which are somewhat lower than the full cohort. Not surprisingly, the 1:3 case-control data sets give estimates closer to the full cohort than the 1:1 sampled data, and for a given number of controls counter-matching gives a slight improvement compared to nested case-control sampling. The differences between the estimates are not big, however, and even the two 1:1 case-control data sets give fairly reliable estimates of the cumulative baseline hazard. This is further illustrated in Table 2 where the first line of each panel (no radon, no smoking) give the increment of the cumulative baseline hazard estimates over the age intervals 40-49, 50-59 and 60-69 years for cohort data and the two 1:1 case-control data sets.

We then computed the cumulative hazard for a given radon exposure history with constant exposure intensity described by the age a at start of exposure, the duration d of exposure, and total exposure. Thus, the two year lagged cumulative radon exposure $R(t)$ is zero for $t < a + 2$, then increases linearly up to the total exposure at $t = a + d + 2$, and is constant at the total exposure thereafter. Smoking was described by the number of packs per day, and we assumed that smoking began at age 20 and continued throughout life at the same level. The increments of such cumulative hazard rate estimates over the age intervals 40-49, 50-59, and 60-69 years are given in Table 2 for cohort data and the two 1:1 case-control data sets with 95% confidence intervals based on the log-transform. The estimates can be interpreted as estimating the absolute

Table 2: Risk (95% confidence interval), in percent, of lung cancer death with specific radon and smoking histories during ages 40-49, 50-59, and 60-69.

Age interval	Total dose* (WLM)	Smoking (packs/day)	Full cohort	1:1 simple	1:1 counter-matched
40-49	0	0	0.24 (0.13-0.44)	0.16 (0.06-0.42)	0.19 (0.07-0.48)
	480	0.5	1.0 (0.7-1.4)	0.8 (0.6-1.2)	1.0 (0.7-1.4)
	960	1.0	2.3 (1.7-3.1)	2.0 (1.4-2.9)	2.4 (1.7-3.3)
50-59	0	0	0.5 (0.3-1.0)	0.4 (0.2-1.0)	0.4 (0.2-1.1)
	480	0.5	3.2 (2.5-3.9)	2.9 (2.3-3.8)	3.0 (2.4-3.9)
	960	1.0	7.9 (6.4-9.6)	7.7 (5.7-10.5)	8.0 (6.2-10.4)
60-69	0	0	0.7 (0.4-1.3)	0.6 (0.2-1.6)	0.6 (0.2-1.5)
	480	0.5	4.5 (3.5-5.9)	5.2 (3.9-7.0)	5.0 (3.8-6.6)
	960	1.0	11.7 (9.2-15.0)	14.3 (10.1-20.2)	13.7 (10.1-18.6)

*Assuming a constant rate of radon exposure for a period of 30 years starting at age 20.

risk of lung cancer deaths in those whom the only reason they would have died during the age interval would be because of lung cancer.[2] There are only small differences in the risk estimates for the two 1:1 sampled data sets, and both of them provide estimates (and standard errors) quite close to what one obtains from the full cohort.

8 Neighborhood-matched counter-matching

Our second example illustrates another, though quite different, application of the idea of counter-matching. We first review the problem which motivated the development of this particular design. Then the design is described in detail, and we discuss how it fits into our general framework. Some power calculations are also reviewed.

8.1 Background

A number of case-control studies have found about twice the rate of childhood leukemia in homes rated as near "very high current configuration" (VHCC) power lines when compared to other power line configurations (Wertheimer and Leeper, 1979; Savitz *et al.*, 1988; London *et al.*, 1991). While each of these studies may be subject to various types of selection bias or information bias, the findings find some support in a population based Swedish study, apparently devoid of such biases (Feychting and Ahlbom, 1995). This finding is especially intriguing given that there is relatively little variation in the rates of childhood leukemia over geography, gender, and ethnicity and that, aside from ionizing radiation, there are no known risk factors. Since VHCCness *per se* cannot cause childhood leukemia, it must be correlated with some factor that does. And, unless very highly correlated, this factor must be much more highly associated with childhood leukemia than VHCC power line configuration. Naturally, electromagnetic fields (EMF) are suspected to be the etiologically relevant exposure, but case-control studies in which extensive EMF measurements were made in the homes of participating subjects have, thus far, failed to

[2]The estimates based on the cumulative hazard estimator give a slight overestimate of risk. However, the difference between these and exact estimates based on a Kaplan-Meier type estimator are of little importance in the situation we consider (Langholz and Borgan, 1997).

show an association with childhood leukemia. One possibility is that power line configuration is a better indicator of long-term EMF exposure than the short-term measurements made in the homes. A second possibility is that EMF is not causing childhood leukemia and that the observed association is due to the correlation of power line configuration with another factor that is etiologically relevant. It is primarily the investigation of the latter possibility that motivates the use of a "neighborhood-matched counter-matched design."

8.2 The design

The study design was developed in collaboration with H. Wachtel and R. Pearson of Radian Corporation, K. Ebi of the Electric Power Research Institute, and D. Thomas of the University of Southern California. We wanted the design to be "highly valid," by which we mean that there should be little chance that the results could be due to selection and information bias. Further, childhood leukemia is a very rare disease so each case should be used "efficiently" in the study. To this end, we wanted to exploit the ability to "wire code" large numbers of homes at relatively low cost using computerized geographic information system methods. In this situation, power-line configuration is the "exposure" and we wish to investigate if there are other factors that can explain its effect. As we discussed in the context of the uranium miners cohort, counter-matching would be an advantageous design for this problem. However, we do not have a well defined cohort, from which to draw our sample. The solution is to define a neighborhood stratified cohort, wire-code those neighborhoods that have cases, and select counter-matched controls based on the within neighborhood risk sets determined by the case. The resulting neighborhood-matched counter-matched study consists of the following steps:

1. Identify incident case through the cancer registry.

2. Wire code a defined neighborhood[3] (VHCC+ vs. VHCC−) surrounding the case's residence at diagnosis. If there is no variation in VHCC status across the entire neighborhood (this may be the situation in some newer developments where all wiring is buried), the case-neighborhood set will be uninformative for this study and so is dropped from the study at this point.

3. For neighborhoods with variation in VHCC status, survey the neighborhood to find all the addresses of all eligible controls (children of a "similar" age, i.e. in the risk set and approximately matched on year of birth).

4. From each neighborhood, form counter-matched sets of one VHCC+ and two VHCC− subjects. That is, if the case is VHCC+ then randomly sample two VHCC− controls and if the case is VHCC−, then randomly sample one VHCC− and one VHCC+ control. The two VHCC−, one VHCC+ configuration was chosen based on power considerations described below.

5. The counter-matched sampled subjects would then be enrolled in the study and information on potential explanatory factors would be collected for the case and its two controls.

Note that the study provides two data sets: (*i*) case-neighborhood data consisting of the cases and their neighborhood controls (selected according to 3), and (*ii*) neighborhood-matched counter-matched data consisting of the cases and their two counter-matched controls (selected

[3]The definition of the neighborhood is a topic worthy of further discussion. But, here, we will assume that some sensible method of forming the neighborhood would be used.

according to 4). Within the framework of cohort sampling, the case-neighborhood data set represents the risk set data from a full cohort with one stratum for each neighborhood. In particular, VHCC status is known (from step 2) for each subject in these risk sets at the failure time so that the full cohort partial likelihood analysis of VHCC is possible (see below). The neighborhood-matched counter-matched data set is sampled from the cohort risk sets as described in Sections 3.2 and 6. So, after collecting potential explanatory factor information in step 5, the counter-matched sampled risk sets consist of one case and two controls, two of the set VHCC− and one VHCC+, with all the information needed for the partial likelihood analysis of VHCC status and potential explanatory factors.

We note that the data sets may be used in different ways depending on the covariate data to be collected. For instance, in the case-neighborhood data set, the information on VHCC status may be supplemented by information on other factors that vary across neighborhoods and are inexpensive to collect for entire neighborhoods (e.g., it may be possible to calculate traffic density measures for each address from existing computerized records). If information on a factor is too expensive to collect for the case-neighborhood set, but does not require actual contact with study subjects (e.g., air and soil samples to assess the presence of environmental pollutants), this may be done on the entire counter-matched sample. Finally, if contact with study subjects is required (e.g., to assess parents' occupations and smoking status), then the counter-matched set would be used but would typically consist of fewer sets because of refusals to participate in the study. Unlike the case-neighborhood and counter-matched sets that do not require subject participation, this latter counter-matched set would be subject to potential biases if these refusals differ systematically between cases and controls with regard to VHCC or other factors of interest. This is true for any study where interviews are required, in particular, for most case-control studies used in epidemiologic research. But, because, the case-neighborhood set defines the entire study group (risk set), it is possible to assess the extent of such a bias, at least with respect to VHCC status.

8.3 Analysis

From the perspective of Section 2, the cohort under study is the entire childhood population serviced by the cancer registry. This cohort is followed for the period of time that the study enrolls cases. This period then determines the entry and exit times for subjects in this amorphous cohort. The underlying hazard model is stratified, as in (22) of Section 6, with a separate baseline hazard for each neighborhood and interval of years of birth. The analysis of relative risk from the case-neighborhood data set, and the neighborhood-matched counter-matched sample use the partial likelihood methods for the full cohort and risk set sampled data, respectively. Since the cohort is highly stratified, the estimation of the stratum specific cumulative hazard is not meaningful, but, further, a pooled estimate, as described in Section 6 is not possible because the number of subjects in the unstratified risk set, i.e. all potential controls in the registry coverage area, is not known.

From the characterization of the case-neighborhood data as the risk sets from the neighborhood (and year of birth) stratified cohort study over the registry coverage area, it is clear that (a stratified version of) the cohort partial likelihood (2) applies. Since neighborhoods with no cases do not contribute to the partial likelihood, only the covariates for the cases and the controls in the surrounding neighborhood are needed for analysis. Further, since childhood leukemia is a rare disease, there will only be one case (a single risk set) from each stratum so that this data is conveniently analyzed using the same conditional logistic software as the sampled data, with weights equal to one.

The neighborhood-matched counter-matched sample is analyzed in the manner described in Sections 3.2, 4 and 6. Since we are proposing to have matched sets with two VHCC− and one VHCC+ subject, we have that $s(i)$ indicates VHCC status for the i subject in the neighborhood, with, say, $n_{VHCC-}(t)$ the number of VHCC− subjects and $n_{VHCC+}(t)$ the number of VHCC positive subjects of age t. Further, the number selected from the sampling strata are $m_{VHCC-} = 2$ and $m_{VHCC+} = 1$. Thus, the appropriate weights $w_i(t)$ (9) for a given set are $n_{VHCC-}(t)/2$ and $n_{VHCC+}(t)$ for the VHCC− subjects and VHCC+ subject, respectively. Again, the weights are determined by the VHCC status only, the case is treated in the same way as controls.

8.4 Power calculations

Although discussion of the asymptotic behavior of estimators from sampled risk set data is beyond the scope of this paper, the asymptotic variance formulas for counter-matched studies (Langholz and Borgan, 1995; Borgan, Goldstein and Langholz, 1995) have an important role in estimating the size of the study needed and the best choice of design parameters. For instance, the decision to use one VHCC+, two VHCC− counter-matched sets was based on a comparison of the power associated with various configurations. Here, we focus on the power of counter-matched sets with this configuration.

The primary goal of these analyses is to identify factors that could explain the association between VHCC-status and childhood leukemia. To this end, we consider the power of the study to detect this association after controlling for the potential explanatory factor. The hypothesis to be tested is whether there is a VHCC association after controlling for another factor given that VHCC is associated when one does not control for the other factor.

A bit more formally, and to explain how the power calculations were done, we start with the "observed relative risk for VHCC+ vs. VHCC−," ϕ, univariately without accounting for any other factors.[4] Now, we want to see if another dichotomous factor Z explains the observed VHCC association. The null hypothesis to be tested is that, after controlling for Z, the relative risk for VHCC is one (i.e. Z explains VHCC). The alternative hypothesis is that, after controlling for Z, the relative risk for VHCC is still ϕ (i.e. Z doesn't explain VHCC at all, either Z is uncorrelated to VHCC or the relative risk for Z after controlling for VHCC is one.)

If Z is to explain VHCC it must be both correlated to VHCC and univariately associated with the disease. We have expressed the correlation between VHCC and Z as the odds ratio

$$\theta = \Pr(\text{VHCC}+, Z+)\Pr(\text{VHCC}-, Z-)/\Pr(\text{VHCC}+, Z-)\Pr(\text{VHCC}-, Z+).$$

We used $\theta = 4$ and 8 for moderate and high correlation between VHCC and presence of the factor $(Z+)$. These may approximately be interpreted as saying that a $Z+$ subject is 4 or 8 times as likely to be VHCC+ than a $Z-$ subject. For a given θ, and assuming that the relative risk for VHCC is one after controlling for Z, the relative risk for Z is determined by the marginal VHCC relative risk ϕ. These are given for various other parameter possibilities in the third column of Table 3. [5]

The powers for counter-matched studies with 200 or 300 cases under some combinations of parameters are given in columns 4 and 5 of Table 3. Assuming that the proportion of VHCC subjects in the neighborhoods is not too low, 200 1:2 counter-matched sets will have sufficient power, 300 sets would be sufficient in the "worst case." For comparison, the powers for a

[4] This corresponds to setting $r(\beta, \mathbf{z}_i) = 1$ for $z_i = VHCC-$ and $r(\beta, \mathbf{z}_i) = \phi$ for $z_i = VHCC+$ in (22).

[5] We note that the relative risks for VHCC of 1.75 and 2.0 bracket are realistic given the results of past studies. The proportion of VHCC positive of 10% is probably quite conservative, given that wire-code homogeneous neighborhoods are discarded so that these power figures are probably low.

Table 3: Power by sample size (N = number of cases), prevalence of the factor that might potentially explain the VHCC effect (Z) (assumed dichotomous) and odds ratio θ between VHCC status and Z. Also given is the relative risk associated with Z if it explains the VHCC association. The proportion VHCC+ in the neighborhoods was taken to be 10%.

Proportion		RR	1:2 Counter-matching*		1:2 Case-control**
Z positive	θ	for Z	$N = 200$	$N = 300$	$N = 300$
Observed relative risk for VHCC $\phi = 1.75$					
10%	4	7.3	67	86	63
10%	8	4.0	66	84	66
25%	4	6.2	71	88	65
25%	8	3.4	70	86	68
Observed relative risk for VHCC $\phi = 2.0$					
10%	4	11.1	85	97	80
10%	8	5.2	85	97	84
25%	4	12.0	88	98	82
25%	8	4.8	86	97	86

Two-sided $\alpha = .05$ level test.

*One VHCC+ subject and two VHCC− subjects.

**The case and two controls randomly sampled from the neighborhood.

case-control study where two controls are randomly sampled from *all* potential controls in the case-neighborhood risk sets are given in column 6 of the table. These powers are about the same as 200 counter-matched sets. Because the rarity of childhood leukemia occurrence[6] is a major limitation for epidemiologic studies of this disease, this increased efficiency greatly reduces the necessary duration time of the study.

A key component both in the cost and validity of the proposed study design is the survey of the neighborhoods in order to locate all potential controls in the neighborhood. The efficacy of "neighborhood walk" methods, in which "walkers" survey a neighborhood by enquiring door-to-door about children who live in the neighborhood (as well as leave letters at homes where this cannot be determined) is currently being investigated under an United States National Institute of Environmental Safety and Health Center pilot project grant. If this method is successful, this design promises to be a useful tool for unraveling the power line-childhood leukemia mystery and, perhaps, will have application in other settings.

9 Concluding comments

The conceptual link between epidemiologic case-control studies and the "study base" (cohort) from which it is drawn has been discussed in many textbooks on epidemiologic methods. However, while this connection is discussed in order to address the potential sources of bias in case-control studies, the link is not invoked in the presentation of the analytic methods for cohort and case-control studies. The risk set sampling approach we have presented here formalizes this connection and unifies the analytic methods, at least with respect to "matched" case-control studies.

Our examples illustrate application of the risk set sampling approach in two very different

[6]For instance, there are about 100 cases per year in Los Angeles County, population 8,000,000.

cohorts. In the uranium miners example, cohort members are individually identified and followed for a long period of time. The risk sets (either calendar period stratified or unstratified) can be exactly set up and sampled. Additional information could then be obtained on this sample. In addition to the estimation of relative risk parameters, because the number at risk is known for each risk set, absolute risks can be estimated from the sample using methods that parallel those for the full cohort. In contrast, in the childhood leukemia example, the cohort is an entire coverage region for a cancer registry and its members are followed for a short period of time. The only cohort members identified are those in the neighborhood-stratified risk sets formed by the cases that occur over the study period. But this is enough information to (counter-match) sample the risk sets and carry out the appropriate partial likelihood estimation of relative risk parameters. Because the cohort is so highly stratified and the numbers in the unstratified risk sets are not known, absolute risk estimation is not possible.

The general analytic framework makes it possible to explore new case-control designs that are adapted to the particular sampling problem. We have illustrated the application of a new procedure, the counter-matching method of case-control (risk set) sampling, in both of our examples. This method exploits information available on the cohort risk sets to obtain a sample that is more informative, with respect to exposure, than random sampling. We have elsewhere described other designs, such as quota sampling (Borgan, Goldstein and Langholz, 1995) and an efficient two-stage case-control study design (Langholz and Goldstein, 1996), that have promise as solutions to the epidemiologic study design problems they were developed to address. We have found that a formal understanding of case-control methodology as sampled risk set data has been tremendously helpful in developing potentially useful new case-control methods.

Acknowledgements

This work was supported by National Cancer Institute grant CA42949 and Electric Power Research Institute contract 4305-02.

References

Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N., and Kreiner, S. (1985). A Cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics*, **41**, 921–932.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100-1120.

Borgan, Ø., Goldstein L., and Langholz (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23**, 1749–1778.

Borgan, Ø. and Langholz, B. (1993). Non-parametric estimation of relative mortality from nested case-control studies. *Biometrics* **49**, 593-602.

Borgan, Ø. and Langholz, B. (1997). Estimation of excess risk from case-control data using Aalen's linear regression model. *Biometrics* **53**, to appear June 1997.

Committee on the Biological Effects of Ionizing Radiation (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters, BEIR IV*, National Academy Press, Washington, D.C.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, 187–220.

Feychting, M. and Ahlbom, A. (1995). Childhood leukemia and residential exposure to weak extremely low frequency magnetic fields. *Environmental Health Perspectives*, **103**, ((Suppl 2)), 59–62.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics* **20**, 1903–1928.

Hornung, R. and Meinhardt, T. (1987). Quantitative risk assessment of lung cancer in U. S. uranium miners. *Health Physics*, **52**, 417–430.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82**, 69-79.

Langholz, B. and Borgan, Ø. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53**, to appear June 1997.

Langholz, B. and Clayton, D. (1994). Sampling strategies in nested case-control studies. *Environmental Health Perspectives* **102 (Suppl 8)**, 47-51.

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science*, **11**, 35-53.

London, S., Thomas, D., Bowman, J., Sobel, E., Cheng, T.-C., and Peters, J. (1991). Exposure to residential electric and magnetic fields and risk of childhood leukemia. *American Journal of Epidemiology*, **134**, 923–937.

Lubin, J., Boice, J., Edling, C., Hornung, R., Howe, G., Kunz, E., Kusiak, R., Morrison, H., Radford, E., Samet, J., Tirmarche, M., Woodward, A., Xiang, Y., and Pierce, D. (1994). Radon and lung cancer risk: A joint analysis of 11 underground miners studies. NIH Publication 94-3644, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Bethesda, MD.

Lundin, F., Wagoner, J., and Archer, V. (1971). Radon daughter exposure and respiratory cancer, quantitative and temporal aspects. Joint Monograph 1, U.S. Public Health Service, Washington, D.C.

Oakes, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *International Statistical Review* **49**, 235–264.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.

Savitz, D., Wachtel, H., Barnes, F., John, E., and Tvrdik, J. (1988). Case-control study of childhood cancer and exposure to 60-hz magnetic fields. *American Journal of Epidemiology*, **128**, 21–38.

Steenland, K. and Deddens, J. A. (1997). Estimating exposure-response trends in nested case-control studies: control selection via counter-matching versus random sampling. *Epidemiology*, in press.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society A* **140**, 469–491.

Thomas, D., Pogoda, J., Langholz, B., and Mack, W. (1994). Temporal modifiers of the radon-smoking interaction. *Health Physics*, **66**, 257–262.

Wertheimer, N. and Leeper, E. (1979). Electrical wiring configurations and childhood cancer. *American Journal of Epidemiology*, **109**, 273–284.

Whittemore, A. and McMillan, A. (1983). Lung cancer mortality among U.S. uranium miners: A reappraisal. *Journal of the National Cancer Institute*, **71**, 489–499.